AD-A050 651 NAVAL HEALTH RESEARCH CENTER SAN DIEGO CALIF F/G 12/1
ARDIE LUBIN ON LINEAR REGRESSION: THE PRODUCT-MOMENT CORRELATIO--ETC(U)
NOV 77 V K THARP

UNCLASSIFIED 77-41 NL

| OF |

AD
A050651

END
DATE
FILMED
4 -78
DDC

12

ARDIE LUBIN ON LINEAR REGRESSION: THE PRODUCT-MOMENT CORRELATION AND THE ONE-PREDICTOR LINEAR REGRESSION MODEL

V. K. THARP, JR.

See 1473 in back

REPORT NO. 77-41

NAVAL HEALTH RESEARCH CENTER
SAN DIEGO, CALIFORNIA 92152

NAVAL MEDICAL RESEARCH AND DEVELOPMENT COMMAND
BETHESDA, MARYLAND

ARDIE LUBIN ON LINEAR REGRESSION: THE PRODUCT-MOMENT CORRELATION AND THE ONE-PREDICTOR LINEAR REGRESSION MODEL

Van K. Tharp, Jr.*

Naval Health Research Center, San Diego, California 92152

KEY WORDS: Correlation, Linear Regression, Least-Squares Model, Statistical Ethics, Diagnostic Checks

This research was supported by the Department of the Navy, Bureau of Medicine and Surgery, under Work Unit MR041.01.03-0152. The views presented in this paper are a summation of the memoranda of the late Ardie Lubin. No endorsement by the Department of the Navy has been given or should be inferred.

*Now at Gateways Hospital, 1891 Effie Street, Los Angeles, CA 90026

PREFACE

Ardie Lubin, shortly before he died, started to teach a course at the Naval Health Research Center on multivariate statistics. The first topic in the course was to be multiple regression, but Ardie discovered that much of the class did not have sufficient background to begin this preliminary topic. Thus, the course consisted of an overview of statistics *à la Lubin* and a thorough coverage of linear regression. Unfortunately, Ardie was admitted to the hospital just as he was ready to begin talking about multiple regression, and the class was never able to get started again.

Ardie produced a series of memos for those first sessions which should be shared with all who can appreciate his giftedness and understanding in the field of statistics. Those memos are the topic of this technical report.

Ardie was in considerable pain when he wrote these memos, which caused him to make several mistakes. I have attempted to correct these errors, and I am to blame if I have not done so adequately. However, the words you are reading are basically Ardie's. To improve the flow of the memo, some of his comments were put as footnotes and I have added some clarification notes of my own.

Ardie promised, at the beginning of one memo, to detail some diagnostic checks. At the end of that memo, the same checks were promised in the next memo. Unfortunately, the "next memo" was never written as Ardie's death occurred on October 9, 1976. Since Ardie talked about the checks in class, I have included them in my own words at the end, along with Ardie's "statement of ethics."

Van K. Tharp, Jr., Ph.D.

Ardie Lubin on Linear Regression: The Product-Moment Correlation and the One-Predictor Linear Regression Model*

The main purpose of this extended note is to supplement the very short discussion of correlation found in Harris' Primer of Multivariate Statistics (1975) on pages 16-19 and 242-3. We will start off with some of the many ways of defining the product-moment correlation and then give a short history to show how correlation and regression came to be linked. Specifically, the relation of the product-moment correlation to the slope coefficient of the usual linear least-squares model will be given. We will point out that what seem to be very slight changes in the specifications of the linear model and the assumptions can completely alter that relation.

DEFINITIONS

The definitions and equations presented here are generally drawn from Kendall and Buckland's A Dictionary of Statistical Terms (1971).

Computation of a product-moment correlation requires two variables, usually denoted X and Y. These variables are quantitative, i.e., at least graded. Many texts call X the independent variable and Y the dependent variable. Kendall and Buckland (1971) point out, however, that this usage is completely incompatible with the standard definition of statistical independence (see p. 71). Thus, psychometricians use the terms predictor and criterion for X and Y, respectively. These terms avoid prejudging the factual issue of dependence.

* Editor's Note: Taken from Ardie's memo dated February 17, 1976.

The linear regression model for a population[a] can be written:

(1) $Y_i = \Psi_o + \Psi_1 X_i + e_i$*

where

Y_i is the criterion score for the ith subject

X_i is the predictor score for the ith subject

Ψ_o is the intercept, the value of Y_i when $X_i = 0$

Ψ_1 is the slope of Y on X, the change in Y when X increases by one unit

e_i refers to any error attributable to the ith subject

When we compute our values from some finite sample of the population, then the linear regression model is:

(2) $Y_i = w_o + w_1 X_i + e_i$

where w_o and w_1 refer to the sample values of the intercept and slope, respectively.

The objective of the linear regression procedure is to find $\hat{Y}_i$, a predicted criterion score for the ith subject, where $\hat{Y}_i$ is equal to the observed score (Y_i) minus the error for that subject (e_i). $\hat{Y}_i$ is the least squares predicted value, that is, it minimizes the squared errors summed over all subjects.**

(a) Population values are usually given Greek letters, while sample values are given Latin letters.

* Editor's Note: A linear equation generally takes the form Y = a + bX, where a refers to the intercept and b to the slope.

** Editor's Note: The discussion of how this is minimized is postponed until a later section.

Let x_i be the deviation of X_i from the mean of the sample (i.e., $x_i = X_i - \overline{X}$) or the deviate score for short and let y_i be the deviate score for Y (i.e., $y_i = Y_i - \overline{Y}$), then the deviance of Y (dev y) is equal to the sum of the squared deviates of Y about the sample mean:

$$(3)\ \text{dev } y = \sum y_i^2 = \sum (Y_i - \overline{Y})^2 \quad \text{(b)}$$

and

$$(4)\ \text{dev } x = \sum x_i^2 = \sum (X_i - \overline{X})^2$$

Most texts talk about the "sum of squares" instead of deviance. I prefer deviance because it is less ambiguous (i.e., it cannot be confused with $\sum X^2$, the sum of the squared raw scores) and less clumsy.* The usual calculation for dev x is:

$$(5)\ \text{dev } x = \sum X_i^2 - (\sum X_i)^2/N$$

Using this notation, the sample variance of X is given by:

$$(6)\ \text{variance } x = \text{dev } x/(N-1) = s_x^2$$

and the standard deviation is the square root of the variance.

Another useful term is that of codeviance. The codeviance of X and Y (i.e., codev xy) is the sum of the cross-products of x_i and y_i, the deviate scores. That is:

$$(7)\ \text{codev } xy = \sum x_i y_i = \sum X_i Y_i - [(\sum X_1)(\sum Y_1)/N]$$

(b) All summations will run from 1 to N, the sample size, unless otherwise noted.

* Editor's Note: I always used the term "sum of squares" until I was indoctrinated by Ardie. The term deviance *is* much less confusing.

Most texts use the term covariance, where:

$$(8)\ \text{covariance xy} = \text{codev xy}/(N-1)$$

Before we discuss the Pearson product-moment correlation and the slope, one other useful concept should be defined, the unit deviate. Unit deviate* scores have a mean of zero and a standard deviation of one. They are found by dividing each deviate score in the sample by the sample standard deviation. In the remaining discussion, let ($u_i = x_i/s_x$) be the unit deviate score for X and let ($v_i = y_i s_y$) be the unit deviate score for Y.

The Pearson product-moment correlation is denoted by r_{xy} for a sample and by ρ_{xy} (pronounced row) for the population. It must be a number between +1 and -1. The product-moment correlation can be defined in a number of ways, utilizing the various concepts presented previously. In terms of unit deviates:

$$(9)\ r_{xy} = (\sum u_i v_i)/N-1$$

when $u_i = v_i$, then $r_{xy} = +1$; and when $u_i = -v_i$, then $r_{xy} = -1$

In terms of codeviance

$$(10)\ r_{xy} = \text{codev xy} \,/\, \sqrt{\text{dev x}}\ \ \sqrt{\text{dev y}}$$

* Editor's Note: A unit deviate score is most commonly referred to as a z-score or a standard score. The reader may find this transformation to be somewhat cumbersome in discussing linear regression. However, the unit deviate transformation is very useful in multivariate statistics, including multiple regression. In multivariate statistics, a set of scores can be represented by a vector. Subtracting the sample mean from each score, i.e., finding deviate scores, locates that vector at the origin of the space in which it is defined. Dividing each score by the sample standard deviation greatly simplifies the computation of the vector's length.

or covariance

$$(11)\ r_{xy} = \text{covar } xy \ / \ \sqrt{\text{var } x} \quad \sqrt{\text{var } y}$$

$$(12)\ r_{xy} = \text{covar } xy \ / \ s_x s_y$$

where s_x and s_y are sample standard deviations.

One additional way of defining the product-moment correlation is very useful for those who, like me, require that a statistic be checked by doing it two different ways. (Remember that redundant information is the only information worth having). Let $d_i = Y_i — X_i$ and $s^2{}_d$ be its variance. Then

$$(13)\ r_{xy} = (s^2{}_x + s^2{}_y - s^2{}_d) \ / \ 2s_x s_y$$

Equation 13 uses the same standard deviations in the denominator as equation 12, but the numerators are completely independent from a computing point of view.

Finally, let me make some points about the relation of r_{xy} to w_1, the least square slope of Y on X in the sample.

$$(14)\ w_1 = r_{xy}(s_y/s_x)$$

or an algebraic derivation

$$(15)\ w_1 = \text{codev } xy \ / \ \text{dev } x$$

Suppose we tried to compute the linear regression weights of equation 2 for u_i and v_i, i.e., suppose we tried to get the least-square prediction of v_i, the unit deviate y score, from u_i, the unit deviate x score. It is easy to show that $w_o = 0$ and $w_1 = r_{xy}$. Thus

(16) $\hat{v}_i = r_{xy} u_i$ and

$\hat{u}_i = r_{xy} v_i$

Thus, the product-moment correlation is also the least-squares regression slope when both variables are in the form of unit deviate scores.

From equation 14, we can see that the equality of the correlation with the slope will always occur when $s_x = s_y$. More fundamentally, r_{xy} is itself a symmetrical concept. The correlation of X with Y is identical to the correlation of Y with X, even when the slope of Y on X differs from the slope of X on Y.

Thus, correlation tends to be used descriptively when we do not distinguish between criterion and predictor or when there is no such distinction. When we can differentiate clearly between the variable we wish to predict and the variable predicted from, the slope tends to be used as a descriptive statistic. Of course, the two statistics are irrevocably linked in that w_1 must be zero if r_{xy} is zero.

Finally, given the sample slope (w_1) and the mean of the criterion and predictor scores (i.e., $\overline{Y}$ and $\overline{X}$, respectively), then it is easy to calculate the intercept, w_o:

(17) $w_o = \overline{Y} - w_1\overline{X}$

HISTORICAL INTRODUCTION[c]

[A] The Two Gaussian Models

As is very common in the history of statistics, we can start by discussing the work of the "Prince of Mathematics," C. F. Gauss. Although Gauss was not

[c] Much of this historical material and discussion is taken from Dudycha & Dudycha (1972) and Binder (1959).

the first to consider the problem of predicting a criterion from a linear function of another variable, his formulation of the problem was used for at least 100 years. Gauss first applied his "Method of Least Squares" to the problem of predicting the orbits of planets. He had to reconcile observations made by different astronomers over several centuries using very different techniques of measurement.

In his first formulation of the problem, Gauss used what we now call the "Method of Maximum Likelihood." Roughly speaking, this means that the predicted value, $\hat{Y}_i$, is chosen so as to be the most likely value for all subjects having the same X score. In order to estimate this modal value of a set of observations, Gauss needed to know the theoretical form of this distribution of the errors of prediction. He took this to be the so-called Normal distribution.[d] Under the assumption that the errors of prediction were distributed normally, the method of least squares was identical to the maximum likelihood method and gave estimates of the regression coefficients with all kinds of optimal properties:

(1) The regression estimates were unbiased. That is, if we average w_1 over all possible samples, that average will equal Ψ_1, the population value.

(2) The w_1 estimates have minimum sampling variance. If we take all the possible sample values of w_1 and calcuiate their variance (i.e., expected mean square) about their average, Ψ_1, then this variance is equal to or less than the sampling variance of any other set of slopes estimated by any other method.

[d] Some statisticians now prefer to call it the Gaussian distribution in view of the contributions made by Gauss to its definition and properties.

(3) The w_1 estimates are distributed normally about Ψ_1, making exact tests of significance possible.

The PLINC Model: The full set of assumptions used by Gauss in this attempt is rather formidable. The first is the most obscure and commonly overlooked:

P All measures of X are perfect with no errors of measurement.

Mathematicians and statisticians usually say that the X values must be mathematical values, whereas the Y values can be expressions of a random variable (i.e., Y_i can possess a stochastic element). Psychometricians, on the other hand, talk about the Xs being infallible measures of the underlying trait. When X is fallible (i.e., has an error of measurement), then the usual least squares estimate of the slope is biased toward zero.

The second assumption is the primary assumption of the regression procedure:

L For each value of X, the mean of the corresponding Y values is a linear transform of X, i.e., Equation 1 holds in the population.

Next is a trio of assumptions that is always made when a t-ratio or F-ratio test is performed:[e]

I The errors of prediction (i.e., the residuals in the model) are independent of one another in the population.

N The errors of prediction are normally distributed about a true value of zero.

[e] Neither test existed in Gauss's day in exact small sample form.

C No matter which value of X is used to predict Y, the variance of the errors of prediction is finite and constant. *That is,* Y is homoscedastic.

Part of the last assumption is sometimes stated separately - the variance of the errors of prediction is finite, never going to infinity. Since this must be true for any finite sample of data, we will take this for granted along with the fact that all data must be quantitative: qualitative variables must be transformed into graded variables for the regression model to apply.

For convenience, I will refer to equation 1 and these five assumptions as the PLINC model. Ordinarily, reference to the linear regression model in a statistics text means the PLINC model, even when all the assumptions are not given explicitly.

The PLI Model: In 1809 Gauss published his method of least squares based on the likelihood approach and these five assumptions. However, after many years of uneasiness with this approach, he reformulated it (1821), eliminating the assumptions that the errors of prediction were normally distributed with constant variance, the PLI model. If the least square equations are solved directly with no reference to these two assumptions, one still gets the same estimators:

$$w_o = \overline{Y} - w_1\overline{X} \quad \text{and} \quad w_1 = \text{codev } xy \ / \ \text{dev } x$$

These estimators are still unbiased and have minimum sampling variance within the class of linear, unbiased estimators.

By eliminating the assumptions of normality and constant variance, the least squares estimator, w_1, will have only the smallest sampling variance of the set of slopes that result from using a <u>linear</u> unbiased estimator. Under the full PLINC assumptions, the sampling variance of w_1 was the smallest possible, no matter whether linear or nonlinear estimates were being compared. Given a non-normal distribution of the errors of prediction and/or a systematic change in their variance as a function of X, some nonlinear unbiased estimators may have a smaller sampling variance than w_o and w_1.

The Central Limit Theorem states that any weighted sum of independent variables tends toward a normal distribution as the number of weighted observations increases indefinitely. By definition, w_1 in the PLI model is a weighted* sum of independent observations. Thus, <u>if enough cases are used, even the test of significance based on the normal distribution will hold</u> for the ordinary least-squares estimators, w_o and w_1, regardless of the true distributions of X and Y.

Even the independence assumption can be modified with no damaging consequences. Suppose that the errors of prediction are correlated instead of independent. If the correlation is constant (i.e., if the correlation between any pair of prediction errors is equal to the correlation between any other pair), then the properties of the sample w_1 values are identical to those for the PLI model except that the variance of w_1 must be decreased to take account of the constant correlation. <u>Constant intercorrelation</u> is sometimes called the <u>compound symmetry condition</u> or the <u>intraclass covariance pattern</u>.

[B] <u>The Contributions of Galton and Pearson</u>

Gauss also generalized his findings immediately to the multiple predictor

* Editor's Note: w_1 is a regression weight

case, and applied "multiple regression" to problems in astronomy, physics, mathematics, etc. Multiple regression, like most good solutions, was rediscovered independently by a number of mathematicians. In addition, mathematicians developed many modifications of the Gauss PLINC and PLI model. In particular, Bravais of France considered what we may call the bivariate-normal model, where Y and X are mutually linear and both marginal distributions are normal.

The idea of a coefficient measuring the goodness-of-fit of a straight line to a set of points originated independently in the fertile mind of Francis Galton (1880). Galton was very much concerned about the regression of English intellect to the moron and idiot level through uncontrolled breeding among the lower classes and undesired immigrants from Europe. Galton was not a mathematician, however, and his idea of a correlation coefficient, varying from +1 to -1 according to the degree of fit, had little impact until Karl Pearson noticed it in Galton's Natural Inheritance (1880). In 1895 Karl Pearson published Contributions to the Mathematical Theory of Evolution[f] in which he derived the large-sample properties of the bivariate-normal correlation and generalized them to the case of multiple correlation.

Galton used the term regression to indicate that parents with extreme measures on height, weight, intelligence, etc. tended to have children for whom these measures were closer to the mean. Under the PLI or PLINC assumptions, such regression to the mean must occur when ρ^2 is not unity.

[f] Both Pearson and Galton believed in the panmixia theory of inheritance: the transmitted elements were in the blood and were continuous in nature, rather than discrete.

Pearson, under the influence of Galton, became very interested in correlation in addition to using the regression model for prediction. In particular, he studied the bivariate case without distinguishing between predictor and criterion. Thus, Pearson's bivariate-normal model was symmetrical, the assumptions for X and Y being the same:

P All measures of X and Y are perfect, with no errors of measurement.

L Y is a linear function of X and vice versa.

I Independence of the errors of prediction.

M Marginal normality: Both X and Y are distributed normally.

C Mutual homoscedasticity. The variance of Y is finite and constant for all values of X and vice versa.

Spearman's Attenuation Factor: Charles Spearman (1904) attacked Pearson on the assumption of perfect measurement. He showed that if random errors of measurement were added to both X and Y, their product-moment correlation could decrease considerably by an attenuation factor. Pearson (1904) did defend his perfect measurement assumption, but I have been unable to find any published acknowledgement by Pearson on the effect of fallible measurement.

Some statisticians (Williams, 1959; Lindley, 1947) now speak of linear functional relations. Williams states:

> "functional relations...subsist between the expected values of different variables and will *not therefore* coincide with regression relations unless the ... variables are free from error ... When the ... variables are errorless, their observed and expected values coincide, so that the regression relation is the same as the functional relation, and both may be estimated by the method of least squares." (Chapter 11)

Like most statistical issues, the correct procedure to use depends on how the question is phrased. If you want to know the correlation between two observed variables or the best least-squares prediction using the observed variables, then the ordinary least-squares procedure is practical and should be used. But if you want to know the correlation between the true variables, then we need to estimate the functional relation between them. Spearman's "correction for attenuation" would be appropriate here, but I have never seen a statistician use it.(g)

Reduction of assumptions. Pearson quickly generalized the bivariate normal model to the multiple correlation case. This meant that the PLIMC assumptions were increased manifold, depending upon the number of predictors. Some of Pearson's co-workers felt that the multitude of assumptions required of the correlation approach swamped its usefulness. Thus, the question of interest became: Could the product-moment correlation be used even *when* all the assumptions did not hold.

G. Udney Yule (1897) discussed the correlation concept with no assumptions except the existence of pairs of numbers. Even under these conditions, some properties hold:

> (1) The correlation must be between +1 and -1. (2) When the standard deviations of X and Y are equal, then the regression of Y on X equals the regression of X on Y which equals the product-moment correlation. (3) When $r^2 = 1$, all points lie on a straight line. (4) As r^2 increases, the sum of the squared deviations from the straight line must decrease. (5) If Y has a non-linear relation to X, then r^2 must be less than unity. (6) The percent of deviance of Y that is accounted for by a linear transform of X is given by r^2.

I find these properties rather disappointing. Unless the observed $r^2 = \pm 1$, I cannot be sure what it means.

Karl Pearson also realized that the PLIMC assumptions were very limiting.

(g) I would have thought that Karl Pearson was interested in the inheritance of the true values rather than the observed ones, but .....

He published (1911) a derivation of multiple correlation which simply used least squares, noting that the computational equations were all the same as when all the multivariate normal assumptions were used. However, he found using the resulting regression weights and correlations for description or inference to be very difficult without assuming at least mutual linearity and marginal normality.

[C] R. A. Fisher and the Exact Small-Sample Approach

Just as Pearson's large-sample approach to the product-moment correlation generated many novel concepts that revolutionized statistics, R. A. Fisher's exact small-sample approach to the sampling distribution of the product-moment correlation was responsible for the next great step in the evolution of present day statistics. To see why the exact small-sample approach shook the statistical world let us compare the Pearson and Fisher approach to the test of the null hypothesis that $\rho = o$.

Pearson's Approach. Pearson would define the estimator of the population parameter to be used for any finite sample. The mean and variance then could be found without reference to the distribution from which the sample was taken. If the population estimator was a linear function of the sample observations, then with a large sample of independent observations the Central Limit Theorem would show that this estimator tended to be distributed normally about the true population mean. For a linear estimator, the population mean and variance are weighted sums of the mean and variance, respectively, of the sample observations. For a non-linear estimator such as the correlation coefficient, Pearson would expand the estimator into a Taylor series and evaluate enough terms to get a usable approximation. Pearson was able to show with such methods that the expected value of the sample r converged to ρ, the population value, as the sample size increased.

The variance of r was a function of only ρ and the sample size.

Pearson's usual test for $\rho = o$ was to divide the sample value by the large-sample approximation to the standard deviation. This ratio was then referred to the normal distribution. In this case, the test was to treat

$$(18)\ z = r\sqrt{N-1}\ /\ \sqrt{1-r^2}$$

as a unit normal deviate. Pearson generally used a very stringent level for significance, z = 3 or more.

Fisher's Approach. R. A. Fisher, who was familiar with Pearson's work, saw a discussion by Soper (1914) on the distribution of the product-moment correlation in a bivariate normal population (i.e., with an infinite N). In a few weeks he sketched out the exact solution and sent it to Pearson. Later, he published a paper (1915) giving the exact distribution of the product-moment correlation for any finite sample from a bivariate normal population (LIMC assumptions).

Fisher's (1915) paper indicated that when $\rho = o$, the expected value (i.e., the average overall possible samples of the same finite size) of the sample product-moment correlation is zero. In other words, when $\rho = o$ in a bivariate normal population, the Pearson product-moment correlation is unbiased.[h] However, this property does not make the Pearson test of the null hypothesis an accurate one with small samples.[i] Fisher recommended using the following ratio:

$$(19)\ t = r\sqrt{N-2}\ /\ \sqrt{1-r^2}$$

[h] Many investigations using all kinds of non-normal distributions uniformly find that the expected value of r is zero when $\rho = o$. The distribution of r is almost normal, being symmetrical about zero with diminishing probabilities as r approaches ± 1.

[i] For Fisher and Pearson a small sample was 100 or less.

Although r is almost normally distributed under the hypothesis that $\rho = o$, the Pearson ratio (equation 18) is not normally distributed and the t-ratio (equation 19) is not normally distributed for small samples. Fisher showed that the t-ratio was generally greater than its corresponding unit normal deviate, but that the difference between the two diminished to near zero as the sample size increased to several hundred cases. A special table, taking the sample size into account, was constructed for the t-ratio.

Thus, the Pearson test of the hypothesis that $\rho = o$ is definitely biased in favor of finding a significant deviation from the null hypothesis.[j] However, the worse case is when ρ is not zero. As ρ deviates towards +1, the distribution of the sample product-moment correlation becomes more and more skewed, with a long tail stretching towards -1. This causes a bias in the expected value of r of approximately:

(20) $-\rho(1-\rho^2) / 2(N-1)$

As ρ deviates towards -1, the bias changes sign.

In summary, r is an unbiased estimate in only three cases: when ρ equals -1, 0, or +1. This is not a particularly welcome conclusion for anyone. As a result, Fisher developed his z transformation of r*, which almost eliminates the bias and almost normalizes the distribution. He recommended the transformation whenever (1) testing if an observed correlation differs significantly

[j] Pearson and his coworkers were very much aware of the fact that their statistical procedures were highly dependent on the use of large samples. In 1947 I had the privilege of attending a class on probability calculus by Florence Nightingale David, a former assistant of K. Pearson. In very brief conversations, I gathered that sample sizes of less than 100 were considered rather incident, if not actually sinful, by workers in the Pearson lab. In a later memo, when we consider the "bouncing beta weights" and the shadowy semi-diaphanous suppressor variables of multiple regression, we shall see the difficulty of justifying the use of small samples in multiple regression.

* Editor's Note: Ardie did not present Fisher's r to z transform. It is given by $z = 1.1513 \log_{10} [(1+r)/(1-r)]$ with a standard error equal to $1/(\sqrt{N-3})$.

from a given theoretical value; (2) testing for a significant difference between two observed correlations; and (3) in combining independent estimates of a correlation to obtain a better one.

Fisher's Approach to the Least-Squares Regression Coefficients. The least-squares regression coefficients, w_o and w_1, are unbiased estimates of the true population values, regardless of the value of the population correlation. Unfortunately, Fisher was unable to derive the sampling distribution and variance of w_o and w_1 for all combinations of N, ρ, σ_x, σ_y and, as far as I know, the exact small sample distribution for all combinations of parameters has never been determined.

Like most mathematicians, Fisher changed the question to one where he knew the answer.[k] Let us go back to equation 15 for the least-squares slope and rewrite it in terms of deviate scores, x_i and y_i:*

$$(21)\ w_1 = \text{codev } xy/\text{dev } x = \sum x_i y_i / \sum x_i^2$$

In other words, each y_i is weighted by $x_i/\sum x_i^2$. Since distribution of a weighted sum of independent normally-distributed variables is well-known to be normal, and the variance of the weighted sum is equal to the sum of the weighted variances, we are practically home free. All that remains is to assume that the y_i values (1) are independent or have the same mutual intercorrelation; (2) are normally distributed for each fixed value of X;

[k] If you can't be near the girl you love, then you love the girl you're near.

* Editor's Note: See page 3 for a discussion of deviate scores.

(3) have constant variance; and (4) that $\overline{Y}$ is a linear function of X. These assumptions of Fisher are frequently called the FLINC model, which is the same as the Gaussian PLINC model except that X is fixed (F).[1] The LINC assumptions apply only to the Y values, just as in the Gaussian model.

The standard error of estimate[m] is a crucial statistic for tests of significance in least-squares linear regression under Fisher's FLINC model. Let us define this statistic in stages. As indicated by equation 2, the observed score, Y_i, is equal to the predicted score, $\hat{Y}_i$ (where $\hat{Y}_i = w_o + w_1X_i$), plus the error of prediction, e_i. If we transform each term to unit deviate form, we get the following equation:

$$(22)\ y_i = w_1x_i + e_i^{*}$$

If we now square each term and sum over all values of i, we then have an expression in terms of deviance.

$$(23)\ \sum y_i^{\,2} = w_1^{\,2}\sum x_i^{\,2} + \sum e_i^{\,2}$$

$$(24)\ \text{dev } y = \text{dev } \hat{y} + \text{dev } e$$

[1] With these assumptions, we have restricted our inferences considerably by making statements of significance and estimation that refer only to future samples having exactly the same distribution of the N values of X. This is known as a conditional test of significance, of which there are many examples in statistics (e.g., all randomization tests, the Chi square test, Hotelling's test of two correlated correlations, etc.).

[m] Also called the standard error of prediction or the standard deviation of the errors of prediction.

[*] Editor's Note: I previously pointed out that transforming scores to their unit deviate form would locate the vector defining those scores at the origin of the space in which the vector is defined. Thus, the intercept, w_o, is eliminated when the scores are transformed to unit deviate form.

That is, the deviance of the observed scores is equal to the deviance of the predicted scores plus the error deviance.

The error variance (n) can now be defined as:

(26) $s_e^2 = \text{dev } e \,/\, (N-2)$

or since dev e = dev y $(1-r^2)$, as

(27) $s_e^2 = \text{dev } y \,(1-r^2)/(N-2)$

To point out the distinction between the criterion y and the predictor x, one can write the error variance as

(28) $s_e^2 = s^2_{y.x}$

where y.x is a convenient short-hand for the prediction error.

Fisher (1915) showed that the best estimate of the sampling variance of w_1, under the FLINC assumptions, is:

(29) $s^2_{w1} = s^2_{y.x} \,/\, \text{dev } x$

(n) A useful outcome of equation 23 is that r^2 is equal to the deviance of the predicted scores divided by the deviance of the observed scores:

(25) $r^2 = \text{dev } \hat{y} \,/\, \text{dev } y$

Since the deviance of the observed scores is the total deviance, the r^2 value is the proportion of the Y deviance predicted by the best fitting linear function of X and $(1-r^2)$ is the proportion of the Y deviance that cannot be predicted from this function. One cannot refer to r^2 as the proportion of the variance of Y that can be linearly predicted from X, because the divisor needed to obtain the variance is not constant. To convert dev y into s_y^2, one must divide by (N-1); whereas to convert dev e into s_e^2, one must divide by (N-2). Of course, for large samples this discrepancy is of no practical importance.

In order to compare the sample slope, w_1, with an hypothesized slope, Ψ_1, Fisher recommended the ratio of the difference to the standard error:

$$(30)\quad t = (w_1 - \Psi_1)\,\sqrt{\text{dev } x}\ /\ s_{y.x}$$

When Ψ_1 is hypothesized to be zero, the t-ratio for the slope is numerically identical to the t-ratio for the product-moment correlation given in equation 19. Thus, when testing whether a least-squares slope is significantly different from zero, the assumption that all future samples will have exactly the same set of X values is not necessary. We can fall back on *the PLINC and LINC* models, depending upon whether we are trying to infer the linear functional relation or simply trying to do a prediction job using the *available measures, respectively.*

W. S. Gosset. Fisher was not the first to work out the exact small-sample distribution of a common statistic. *This honor goes* to W. S. Gosset, a student of Karl Pearson, who worked as a chemist for the Guiness brewery (and presumably was allowed *only small samples of the brew!*). Gosset wrote under the name of Student to preserve commercial security.

In 1908 Student published a paper giving the exact small sample distribution of the ratio:

$$(31)\quad z = (\overline{X} - \mu)/s_x$$

where $\overline{X}$ is the sample mean, μ is the population mean, and s_x is the sample standard deviation with (N-1) as the denominator.

Fisher did not know of Student's (1908) paper when he published his 1915 paper. Later, he gave full credit to Student and declared that one of the chief purposes of Statistical Methods for Research Workers was to make the work of Student appreciated and better known. Student was unable to prove that the distribution of the standard deviation he was using was the correct one, so later Fisher (1925) gave a rigorous proof of Student's results. Fisher also invented the actual form of the t-ratio, since Student used the z-ratio in his 1908 paper (equation 31) which does not generalize quite as easily as the t-ratio.

Fisher's (1915) paper was received much like Student's (1908) paper -- with deafening apathy. Fisher spent the next four years teaching physics and mathematics. About 1919 he was, almost simultaneously, offered a post with K. Pearson and a position as a statistician at the Rothamsted Agricultural Station. He chose Rothamsted. Judging from what we now know of the two personalities involved, he would not have remained long with K. Pearson.

[D] The Pitman Permutation Test

As happens so frequently in statistics, the permutation test started with the work of R. A. Fisher. Fisher (1925, section 21) was attempting to answer the criticism that the Student t-test is seriously limited because it requires that the observations be drawn from a normal distribution. He asked whether a materially different result would be obtained if one assumes that all observations are independently drawn from the same population (i.e., the homerous assumption) without specifying a normal distribution for this common population.

The example being considered was a test of the difference between correlated means with 15 pairs of observations. If each (X,Y) pair was truly drawn at random from the two series, then X-Y would have occurred as frequently as Y-X. In all, some $2^{15} = 32,768$ average differences could be generated by such random combinations. The observed average difference of 20.933 was exceeded, in the positive or negative direction from zero, by 5.267% of the 32,768 chance combinations. The t-ratio was 2.148, which gives a two-tailed level of 4.97% by the normal based table. Thus, Fisher argued that the requirement of normality is not a serious limitation, at least in this case, since almost the same significance level can be obtained by just assuming independent data from the same population.

Other statisticians saw almost immediately that the randomization or permutation method could be used to make a conditional test of the null hypothesis for any statistic where the t-ratio or F-ratio would ordinarily be used. In particular, E. J. G. Pitman (1937) applied the permutation method to the product-moment correlation to test the null hypothesis that $\rho = o$.

In principle, the test is simple. We calculate the value of the observed correlation in the usual way. Then we permute the order of the Y values and recalculate the correlation. This procedure is repeated until a product-moment correlation is obtained for each possible permutation of the order of the Y values. For N pairs of observations, there are N! permutations and thus N! product-moment correlations. These N! correlations correspond to the sampling distribution of coefficients under the randomization null hypothesis: That is, the sample was drawn from a population in which the X,Y pairs were formed by chance. The N! correlations are arranged in order of magnitude and the percent of coefficients greater than the observed product-moment correlation is calculated. When the percentage of permutation coefficients exceeding the observed product-moment correlation is less than 5%, one can reject the null hypothesis at the 5% level.*

Table 1

Exercise on Randomization Test

Y	a	b	c	d	e	f	g	h
0	0	0	0	0	0	0	1	1
2	1	1	2	2	9	9	0	0
5	2	9	1	9	1	2	2	9
9	9	2	9	1	2	1	9	2
r_{xy}	.938	.355	.876	.209	-.146	-.229	.897	.313
w_{yx}	.90	.34	.84	.20	-.14	-.22	.86	.30
w_o	1.30	2.98	1.48	3.40	4.42	4.66	1.42	3.10

Y	i	j	k	l	m	n	o	p
0	1	1	1	1	2	2	2	2
2	2	2	9	9	0	0	1	1
5	0	9	0	2	1	9	0	9
9	9	0	2	0	9	1	9	0
r_{xy}	.772	.021	-.250	-.417	.792	.125	.730	-.021
w_{yx}	.74	.02	-.24	-.40	.76	.12	.70	-.02
w_o	1.78	3.94	4.72	5.20	1.72	3.64	1.90	4.06

Y	q	r	s	t	u	v	w	x
0	2	2	9	9	9	9	9	9
2	9	9	0	0	1	1	2	2
5	0	1	1	2	0	2	0	1
9	1	0	2	1	2	0	1	0
r_{xy}	-.438	-.521	-.521	-.605	-.584	-.751	-.730	-.813
w_{yx}	-.42	-.50	-.50	-.58	-.56	-.72	-.70	-.78
w_o	5.26	5.50	5.50	5.74	5.68	6.16	6.10	6.34

* Editor's Note: Ardie handed out an exercise on the randomization test for the product-moment correlation. This exercise is included as Table 1 as an illustration for the reader.

Note the highly *conditional* nature of the permutation test. Strictly speaking, the significance of the observed correlation is being judged only with reference to a population having marginal distributions of X and Y that are identical to the sample. Thus, any inference about the value of the population correlation holds only for future samples with exactly those marginal distributions of X and Y. Of course, few researchers are very strict about their inferences when they publish such results, but it is salutary to realize that the marginal distributions of X and Y for the drawn sample may not represent your population. If so, then your inference about the population correlation may be sadly askew.

Pitman was able to prove some generalizations about the moments of the permutation distribution of correlations under the null hypothesis. Perhaps his most important conclusion was that as N increases indefinitely, the distribution of the permutation correlations tends toward the distribution of the bivariate normal product-moment r. Thus, the Pitman permutation test for a large N is identical to the Fisher t test for the null hypothesis that $\rho = o$ (i.e., equation 19).

The Pitman results are worth presenting in some detail. *The average* of all N! permutation correlations *will always be exactly zero, and the variance* of the N! coefficients about zero *is exactly equal to the inverse* of (N-1). The third moment of the distribution, μ_{3r}, is:

$$(32)\quad \mu_{3r} = \sum r^3/N! = (N-2)\gamma_{x1}\gamma_{yi}/N(N-1)^2$$

where γ_{x1} and γ_{y1} refer to the skewness of the X and Y distributions, respectively. The general equation for skewness is:

$$(33)\quad \gamma_{x1} = E\ (x_i/\sigma_x)^2$$

where E denotes expectation over all possible values of X, and x_i represents the deviation from the population mean of X, μ_X. Thus, γ_1 will be positive if there is a long positive tail; negative if there is a long negative tail; and zero if the distribution is symmetric about the mean, as in the case of a normal distribution.

Equation 32 implies that if both X and Y have skewed distributions, then some extremely high values of r^2 will occur in the permutation distribution. If both variables are skewed either negatively or positively, then the distribution of permutation correlations will have some extremely large positive correlations when the skewed observations from X and Y happen to coincide. When one variable has a positive skew and the other a negative skew, then the distribution will contain some correlations close to -1. When one of the variables has no skew, then the problem of bivariate skew vanishes and the distribution of the permutation correlations will have no skew. Finally, the most important implication of equation 32 is that the skew of the permutation correlation distribution must approach zero as N increases.

The standardized fourth moment (i.e., the Kurtosis)of the permutation correlation distribution will equal 3/(N-1)(N+1) if either X or Y has a Kurtosis of zero. Moreover, as N increases the Kurtosis of the permutation correlation distribution approaches 3/(N-1)(N+1), irregardless of the Kurtosis of either X or Y.

The importance of all this is that Fisher's test of significance for the product-moment correlation (equation 19) can be used. Thus, using only the assumptions of independence of the N pairs and mutual linearity of X and Y, we are able to use the same test of significance as the bivariate normal model with the LIMC assumptions.

What is the moral of the Pitman story? First, if you simply concentrate on one question, whether the null hypothesis is true in the population, a conditional test of significance is always possible with very few assumptions. Second, the large sample version of the permutation test of significance may (and generally does) coincide with the exact test of significance used when all the bivariate normal assumptions hold in population.

There are two disadvantages to the Pitman permutation test. First, the test is conditional and thus limited to the population where the marginal distributions of X and Y are identical to those observed in the sample. Second, by the time the sample size is large enough for one to expect a stable correlation (i.e., N=20), the number of permutation correlations required to obtain the distribution has soared to the millions. The only saving grace here is that if either X or Y has a symmetric distribution (i.e., a skew of zero) and is very flat (i.e., the Kurtosis is near zero), then the usual t-test is valid for small samples without the necessity of constructing the entire permutation correlation distribution.

[E] <u>The Spearman Correlation for Ranked Observations</u>

C. Spearman in his article (1904) raising the issue of errors of measurement and their attenuation effect on the product-moment correlation, also suggested that each variable be ranked from 1 to N and a product-moment correlation calculated for the ranks. Pearson (1907) responded favorably to this latter suggestion and worked extensively on the characteristics of rank correlations.[o]

One immediate question was how to determine the significance of the Spearman rank correlation. First, when $\rho = o$, the variance of the sample value is 1/(N-1). This is identical to Pitman's (1937) result for the permutation product-moment r distribution. However, like the product-moment correlation, the Spearman rank correlation has distinctly non-normal distributions for small samples or when ρ is near ± 1. Thus, the significance of an observed rank correlation (hereafter designated sr) cannot be tested by dividing it by its variance.

In 1936, H. Hotelling and M. Pabst used the permutation method of generating the sampling distribution of the sr in the null case that $\rho = o$.[p] They managed to work out the exact distribution of sr for N=2 to N=7. For N=7 they had to calculate 7! = 5040 values with no electronic computer to help them, so they stopped there. However, Kendall, Kendall, and Babington (1938) soon after computed the 8! = 40,320 values of sr for N=8; and David, Kendall, and Stuart (1951) later published the exact probability levels for N=9 and N=10. As far as I know, no one has worked out the 11! = 39,916,800 Spearman rank correlations for N=11, but there is no practical need to do so. The Fisher t-test for the significance of a bivariate normal product-moment correlation will give almost exactly the correct levels of significance for sr with N greater than 10. In other words, all that is necessary is to substitute sr for r in equation 19.

[o] Incidentally, Spearman and Pearson spent much of their working lives in close physical proximity at University College, London. Spearman was head of the Psychology Department and Pearson was head of the Biometrics and Eugenics Laboratories, so the two men were separated only by a courtyard. Apparently, they had no particular liking for one another. Their disputes stimulated the development of some important concepts and led Pearson to exhaustive mathematical investigations for which Spearman had no training.

[p] The use of the permutation method was apparently quite independent of Fisher or Pitman.

Table 2 illustrates the degree of approximation involved. Here I have listed the .05 one-tail values of the Pearson product-moment correlation based on the PLIMC assumptions of the bivariate-normal model and the transformation to the Fisher t-ratio. Since the Spearman rank correlation can only take on a finite set of discrete values, most of the sr$_{.05}$ values given in Table 2 are fiction in the sense that no such sr values could ever result for the given sample size of N. The sr$_{.05}$ values were obtained by linear interpolation from the exact probabilities and tend to be a bit too high.

Table 2

Values of the Spearman Rank-Order and Pearson Product-Moment Correlations at the .05 one-tailed Level of Significance

N	df=(N-2)	r$_{.05}$	sr$_{.05}$
4	2	.900	.987
5	3	.805	.908
6	4	.729	.774
7	5	.669	.695
8	6	.621	.637
9	7	.582	.583
10	8	.549	.546

Table 2 shows that as N increases, the two critical .05 levels approach one another, until, for N greater than 10, they differ only in the third decimal place. Practically all sample sizes above 10 are large samples when testing sr for significance at the .05 one-tail level.

Hotelling, Pabst, and Kendall generated the sampling distribution of the Spearman rank correlation under the null hypothesis exactly in the way called for by the Pitman permutation method. Therefore, the moments of the sampling distribution of sr, under the null hypothesis, can be deduced from the general equations given by Pitman. The average of all N! rank correlations will be zero and the variance will be 1/(N-1). For the Spearman rank correlation both X and Y have been transformed to rectangular distributions, which are symmetric about (N+1)/2. Thus, the skew coefficients for X and Y are zero and so is the sampling distribution of the Spearman rank correlation under the null hypothesis. The fourth moment is given by:

$$(34)\ \mu_{4sr} = 3K/(N^2-1)$$

where K = 12(N+6)/25N (N-1). As N increases, K goes to zero.

The sampling distribution of sr is identical to the Pearson Type II symmetrical distribution, except for the fourth moment which becomes similar as N increases. Fisher (1915) proved that the exact distribution of the product-moment correlation, under the LIMC assumptions with $\rho = o$, is the Pearson Type II symmetrical distribution. Therefore, to the extent that the moments of the Spearman rank correlation sampling distribution approach those of the Pearson Type II distribution, the exact Fisher t-test (equation 19) will be a good approximate test for the significance of sr.

I view the Spearman rank correlation as an excellent alternative to the product-moment correlation when the question is whether there is a significant association between X and Y. When you have drawn the subjects independently and assured yourself of a mutual linear regression, then transforming X and Y into ranks gets rid of the assumptions of marginal normality and homoscedasticity.

However, rank correlation is not an answer to the prediction of the observed Y scores from the observed X scores with the least error. For the same number of subjects, sr will have less statistical power (i.e., the probability that the experimenter will accept the alternative hypothesis) than r. Hotelling and Pabst (1936) found that for large samples r attained the same power as sr with 91% of the sample size. Thus, the asymptotic efficiency of sr is 91% of that of r when the bivariate normal assumptions are met. Of course, sr may be more efficient than r when the bivariate normal assumptions are not met. In fact, sr is the most efficient measure of correlation when the marginal distributions of X and Y are logistic functions.

ADDENDUM: DIAGNOSTIC CHECKS

In my opinion, the right to compute any statistic you wish, even a product-moment correlation, is guaranteed by the first Article of the Bill of Rights just as much as is freedom of speech or religion. But just as freedom of speech is not an adequate response to the charge of libel or publishing false and misleading commercial information, just so the right to compute a statistic is not the right to publish false or misleading information. Just as the judges and the courts stand guard over the rights of citizens and try to prevent abuse of freedom of speech, just so editors and referees stand guard over the rights of the reader and try to prevent publication of statistics that misrepresent the data, contain spurious elements, or are just plain false.

Ordinarily then, the question of whether a statistic should be computed is not an ethical question for me. If you think it would yield information of interest to you, then of course the answer is 'Hell, yes, compute it.' But there is an ethical violation, for me, when someone attempts to publish a statistic before (1) checking for numerical accuracy, (2) checking that essential parts of the model do fit the data, (3) checking that artifact or spurious elements have not distorted the numerical value of the statistic, (4) making sure that the published level of significance has not been distorted by multiple comparisons, etc.

Moral standards change from time to time in statistics as well as in social fields. At present, very few would agree that diagnostic checks, of how well the model fits the data, are just as much a responsibility of the research worker as numerical checks of accuracy. To my surprise, some editors apparently don't even agree on the necessity for numerical checks of accuracy. Inevitably, one can point to wildly inaccurate statistics being published in such journals. I would hope in the near future, that editors will insist on statements from the authors regarding such checks just as they now insist on statements regarding the humane treatment of subjects.

Ardie recommended the following steps as diagnostic checks for the linear regression model for the one-predictor case.

(1) Test for linearity:

$$\text{Compute } F = (r_q^2 - r_{xy}^2)/\frac{(1-r_q^2)}{(N-3)}$$

where r_{xy}^2 is the Pearson product-moment correlation and r_q^2 is the quadratic correlation coefficient.

or

$$\text{Compute } F = \frac{(CR-r_{xy}2)}{(K-2)} / \frac{(1-CR)}{(N-K)}$$

where CR is the correlation coefficient and r_{xy}^2 is the Pearson product-moment correlation.

If either of these are nonsignificant, go to step 4. Otherwise, go to step 2.

(2) Test for quadratic model:

$$\text{Compute: } F = \frac{(CR - r_q^{\ 2})}{(K-3)} \;/\; \frac{(1-CR)}{(N-K)}$$

If significant a nonquadratic, nonlinear model describes the data. If nonsignificant, the quadratic model is sufficient.

(3) Transform the data to obtain linearity and repeat steps 1 and 2.

(4) Test for skewness:

Compare the Pearson product-moment correlation with the Spearman rank order correlation. These results should agree with step 1.

(5) Plot the data. These results should agree with those obtained in steps 1 and 2.

REFERENCES

Binder, A. Considerations of the place of assumptions in correlational analysis. American Psychologist, 1959, 14, 504-510.

David, S. T., Kendall, M. G., & Stuart, A. Some questions of distribution in the theory of rank correlation. Biometrika, 1951, 38, 131-140.

Dudycha, A. L., & Dudycha, L. W. Behavioral Statistics. In R. E. Kirk (Ed.), Statistical Issues. Monterey, CA: Brooks-Cole Pub., 1972, pp. 2-25.

Fisher, R. A. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika, 1915, 10, 507-521.

Fisher, R. A. Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh, 1925.

Galton, F. Co-relations and their measurement chiefly from anthropometric data. 1880.

Gauss, C. F. Carl Friedrich Gauss Werke. (collected works, 12 vol.) Gottingen: Dieterichsche Universitats-Druckerei, 1870-1933.

Harris, R. J. Primer of multivariate statistics. New York: Academic Press, 1975.

Hotelling, H., & Pabst, M. Rank correlation and tests of significance involving no assumption of normality. Annals of Mathematical Statistics, 1936, 7, 29-43.

Kendall, M. G., & Buckland, W. R. A dictionary of statistical terms. New York: Hafner Pub., 1971.

Kendall, M. G., Kendall, S. F. H., & Babington, B. The distribution of Spearman's coefficient of rank correlation in a universe in which all rankings occur an equal number of times. Biometrika, 1938, 30, 251-273.

Lindley, D. V. Regression lines and the linear functional relation. Journal Royal Statistical Society Supplement, 1947, 9, 218-244.

Pearson, K. On the inheritance of the mental and moral characters in man and its comparison with the inheritance of physical characters. Biometrika, 1904, 3, 131-160.

Pearson, K. On the general theory of the influence of selection on correlation and variation. Biometrika, 1911, 8, 437-443.

Pitman, E. J. G. Significance tests which may be applied to samples from any population. II. The correlation coefficient. Supplement to the Journal of the Royal Statistical Society, 1937, 4, 225.

Spearman, C. The proof and measurement of association between two things. American Journal of Psychology, 1904, 15, 72-101.

Student. The probable error of the mean. Biometrika, 1908, 6, 1-25.

Williams, E. J. Regression analysis. New York: Wiley, 1959.

Yule, George Udney. On the theory of correlation. Journal Royal Statistical Society, 1897, 60, 812-854.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (*When Data Entered*)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 14 77-41	2. GOVT ACCESSION NO.
	3. RECIPIENT'S CATALOG NUMBER 9
4. TITLE (*and Subtitle*) 6 Ardie Lubin on Linear Regression: The Product-Moment Correlation and the One-Predictor Linear Regression Model.	5. ~~TYPE OF REPORT & PERI~~OD COVERED Interim rept.,
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(*s*) 10 Van K./Tharp, Jr.	8. CONTRACT OR GRANT NUMBER(*s*) 16 MRØ41Ø1 17 MRØ41Ø1Ø3
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Health Research Center San Diego, CA 92152	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS MR041.01.03-0152 61153N
11. CONTROLLING OFFICE NAME AND ADDRESS Naval Medical Research & Development Command Bethesda, MD 20014	11 12. REPORT DATE Nov[redacted]77
	13. NUMBER OF PAGES 34 12 38p.
14. MONITORING AGENCY NAME & ADDRESS(*if different from Controlling Office*) Bureau of Medicine and Surgery Department of the Navy Washington, D. C. 20372	15. SECURITY CLASS. (*of this report*) Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (*of this Report*)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (*of the abstract entered in Block 20, if different from Report*)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (*Continue on reverse side if necessary and identify by block number*)

Correlation, Linear Regression, Least-Squares Model, Statistical Ethics, Diagnostic Checks.

20. ABSTRACT (*Continue on reverse side if necessary and identify by block number*)

(U)...This Center Technical Report is divided into three parts. A series of definitions and formulas necessary for understanding and computing the product-moment correlation are presented in Part I. In Part II, the historical development of the linear regression model is presented, illustrating various models and their assumptions. Finally, Part III contains Ardie Lubin's statement of ethics for using statistics in research and a series of diagnostic checks for the one-predictor linear regression model.

391 642